Sai Sri Harsha Guddati

Full-stack software engineer | Real-time systems and AI copilots

Gainesville, FL | (352) 665-8709 | saisriharshaguddati1@gmail.com | Portfolio | GitHub | LinkedIn

EXPERIENCE

University of Florida Information Technology May 2025 - Present

Current role: Infrastructure Platform Engineer IV Gainesville, FL

- Led development of **NaviGator Prompt Builder** across Next.js/TypeScript and FastAPI, connecting the prompt creation, evaluation and sharing interface to backend AI services for university users.
- Redesigned real-time delivery with Redis Pub/Sub, Celery and Server-Sent Events for **2,000+ concurrent token streams**, reducing p95 response latency from **1.2 s to 450 ms** across 12+ generative AI workflows.
- Designed document pipelines for PDFs, policies and institutional records with OCR fallback, chunking, embeddings and metadata filters. Combined hybrid retrieval, reranking and source attribution to ground answers in the underlying documents.
- Built evaluation datasets and regression harnesses for retrieval relevance, groundedness, tool-call correctness, JSON-schema compliance, latency and provider-specific failures. Used results to tune prompts, model routing and output validation.
- Maintain and extend UFIT's LiteLLM-based gateway on a shared platform processing **9B+ tokens/day**. Configure provider routing and failover, API-key authentication, RBAC, team isolation, budgets and audit logging for multi-tenant LLM access.
- Deploy and operate **10+ containerized microservices** on Kubernetes with Terraform-managed infrastructure, F5 load balancing and GitLab CI/CD. Maintain automated tests, health checks and staged rollouts for production releases.
- Built a departmental Data Science Agent that ingests Excel/PDF sources, including datasets with hundreds of thousands of rows. Combined TimesFM forecasting with BM25 keyword search and dense-vector retrieval over PostgreSQL/pgvector to answer data questions.

Tata Consultancy Services Aug 2022 - Dec 2023

Systems Engineer Hyderabad, India

- Built Java Spring Boot and Node.js services supporting **100K+ daily user interactions**, with request validation, idempotent processing and backward-compatible REST contracts for independently deployed frontend and downstream systems.
- Reduced average API response time by **approximately 50%** across high-volume workflows by tuning PostgreSQL/MSSQL query plans, joins, indexes and connection usage and refactoring inefficient data access.
- Moved high-volume ingestion out of user-facing request threads into asynchronous workers with retries and rate limits. Maintained GitLab CI/CD pipelines for automated testing, packaging and multi-environment deployments.

SELECTED PROJECTS

WebLMS | TypeScript, WebExtensions, FastAPI, Redis | GitHub

- Developed a browser assistant that extracts page context through DOM parsing and OCR, applies client-side privacy filters, and coordinates Redis caching, background tasks and streamed responses for contextual AI chat.

IntentGuard | Python, LiteLLM | PyPI

- Published an alpha Python command-line agent that intercepts commands, gathers execution context and evaluates risk through LiteLLM. Explains risky actions before execution and lets users confirm or abort them.

TECHNICAL SKILLS

Languages and web: Python, TypeScript, Java, SQL, React, Next.js, FastAPI, Spring Boot, Node.js

Data and AI: PostgreSQL, MSSQL, pgvector, Redis, Celery, LiteLLM, RAG, model evaluation

Infrastructure: Kubernetes, Docker, Terraform, GitLab CI/CD, Grafana, Splunk, JWT, SAML/SSO, RBAC

EDUCATION

University of Florida | M.S. in Computer Science

Jawaharlal Nehru Technological University Kakinada | B.Tech in Computer Science